Genome wide identification of regulatory networks associated with general cognitive ability using a normalized alignment free similarity measure of promoter regions

Miriam Ruth Kantorovitz[1,*], David Tcheng[2], Michael I. Lerman[3], Eric Jakobsson[4]
1 Department of Mathematics University of Illinois, Urbana, IL, USA
2 National Center for Supercomputing Applications University of Illinois, Urbana, IL, USA 3 Scientific Board, Affina Biotechnologies, Inc. Stamford, CT, USA
4 Department of Molecular and Integrative Physiology, Department of Biochemistry, UIUC programs in Biophysics, Neuroscience, and Bioengineering, National Center for Supercomputing Applications, and Beckman Institute, University of Illinois, Urbana, IL, USA
∗ **E-mail: ruth@math.uiuc.edu**

Abstract

We show that a normalized alignment free similarity measure, called D2z, can be used to detect potential regulatory relations for gene sets when little is known about the regulatory elements involved. One scenario where such gene sets arise is genome wide association studies (GWAS). In this work we consider a gene set from a GWAS on childhood general cognitive ability. We build a co-regulation network for the GWAS genes based on the D2z scores, which shows potential co-regulatory relationships between the genes as well as predict additional genes that are likely to be part of the network. We found that the set of the predicted genes is enriched in genes associated with mental retardation and GO terms such as synapse and neuron development. In particular, we found strong evidence of regulatory connection between the GWAS genes and CHL1, a gene known to be involved in mental retardation.

Introduction

The genetic architecture of childhood general cognitive ability (g) is complex. Many genes with small effects may be involved rather than a few genes with large effects ([1] and references within). In a large genome-wide association study (GWAS), nine SNPs were found to be associated with g [1]. These loci were mapped to eight genes with non-overlapping promoter regions. In this paper we use a normalized alignment free similarity measures, called D2z, applied to promoter regions of genes, to predict co-regulation relationship between these genes associated with g.

The similarity of two biological sequences has traditionally been assessed within the well-established framework of alignment. However, when studying gene regulation, one often needs to identify functional relationships between DNA sequences that do not exhibit any statistically significant alignment. This is the case, for example, when comparing cis-regulatory modules (CRMs) in the promoter regions of non-orthologous genes. The alignment free similarity measure D2z [2, 3] was shown to detect functional similarity between regulatory sequences. In [2] we showed that the D2z method can accurately discriminate functionally related CRMs from unrelated sequence pairs in fruit fly and human. In [3] the D2z measure was used for genome-wide searches in fruit fly and human to predict CRMs with a similar function to a given set of CRMs known to mediate a common gene expression pattern, but without the use of known motifs. These predictions had been validated successfully in vivo. Various alignment-free similarity measures had been proposed previously (e.g., [4] and references within), however, when applied to real biological data, these methods were unable to discriminate functionally related CRMs from unrelated sequence pairs [2]. The main problem with these methods is that they are not normalized. That is, the similarity scores are not comparable across sequence pairs drawn from arbitrary background

distributions. More recently, other promising normalized alignment free methods have been developed, such as [5, 6].

In this work we test whether the D2z measure, when applied to promoter regions of genes, can be used for detecting possible regulatory relations between genes, without using prior knowledge of the transcription factors (TFs) that may be involved. This is the case for the set of genes implicated in general cognitive ability in the GWAS, where the interactions or co-regulation relationship between the genes is not known. As in the CRM case, the idea is that if two genes have similar promoter regions, then these promoter regions are likely to contain binding sites for the same (unknown) TFs that target these genes and hence the two genes are likely to be co-regulated. However, unlike CRMs, which usually contain multiple binding sites for relevant TFs, a promoter region may not be as enriched in TF binding sites, and therefore the similarity is harder to detect. We show that our method detects co-regulatory relationship between the eight genes associated with g, as well as connects them to genes that are known to be involved in brain development or intellectual disabilities.

From previous work [7–9], we were particularly interested in regulatory connections between the gene Close homologue of L1 (CHL1, also known as CALL) and the cognition genes from the GWAS study. CHL1 is a neural recognition molecule that plays important roles in cell migration, axonal growth, and synaptic remodeling. It has been associated with mental retardation, reduced intelligence, schizophrenia, memory and behavior disorders [7–16]. Our method showed a strong connection between CHL1 and the cognition genes. We perform a genome-wide search for additional genes that are potentially co-regulated with the genes found in the GWAS. We found that the set of the predicted genes is enriched in genes associated with mental retardation and GO terms such as synapse, exonogenesis and neuron development.

The advantage of our approach is that it does not rely on TF databases and does not use prior knowledge of the TFs that may be involved in the network. Since the only input we use is the genes promoter regions, this method can be readily applied to any genome. Other normalized alignment free methods (e.g. [5, 6]) could potentially be used in the same manner. This approach, which also does not use information from the coding regions, can be combined with methods that use such information, to better understand the networks in question.

Results and Discussion

In the GWAS study [1], eight genes with non-overlapping promoter regions were associated with general cognitive ability (g), which were also in the DBTSS database: RXRA, CARS, CTNNA3, TMCC3, MAP3K7, STK10, NR2F1 and FERMT1. We first construct the co-regulation graph for these genes as follows.

Constructing the co-regulation graph

For a given set of genes, we construct a co-regulation graph using the promoter region of the genes (see Methods for details). Briefly, each promoter region from the dataset ("probe") is scored against the promoter regions of all other human genes in the database DBTSS [17] using the D2z similarity measure [2]. For each probe gene, the genes associated with the top scoring promoter sequences are retrieved. This relationship is captured by a directed edge in the co-regulation graph, from the probe gene to each of the top scoring genes. Thus, the edges in the co-regulation graph represent a co-regulation relationship. A highly connected graph means that the probe genes have many commonly co-regulated genes and therefore these genes are likely to be part of a regulatory network. In this work we used the top 30 scoring genes for each probe, which is about 0.1 percent of the sequences in the DBTSS database.

Figure 1 shows the co-regulation graphs for these genes. We see that 7 of the 8 probes are connected, with 25 shared nodes ($p < 10^{-5}$). One gene, CARS, is not connected to the rest of the graph.

Relations with CHL1 and genome-wide discoveries

To find potential co-regulatory relationship between CHL1, which is a gene known to be involved in mental retardation, schizophrenia, memory, and social behavior [7–16], and the GWAS cognition genes, we constructed the co-regulation graph for CHL1 with the eight cognition genes (Figure 2). Interestingly, when we added CHL1 to the probe gene set, the co-regulation graph became connected with 33 shared nodes ($p < 10^{-5}$). In addition, CHL1 had a shared node with seven of the eight probes from the GWAS study, and a direct edge to one of the probes: NR2F1. Among the genes that CHL1 and NR2F1 share in the co-regulation graph (that is, genes that are commonly co-regulated with both, CHL1 and NR2F1) were genes that are known to be involved in learning disabilities and brain development, such as, SYNGAP1 [18, 19] and NFIX [20]. A complete list of the top 30 scoring genes for each probe is provided in supplementary material. This list includes genes that are potentially co-regulated with CHL1 but are not in the co-regulation graph since they were not commonly co-regulated with genes from the GWAS data set.

To discover more genes with potential co-regulatory properties as CHL1 with respect to the GWAS gene set, we searched genom-wide for genes that are commonly co-regulated with at least one of the genes in the GWAS gene set (i.e., genes with a direct edge to the probe set in the co-regulation graph). We found about 900 such genes, which are potentially part of the regulatory network involved in general cognitive ability (see supplementary material).

Using the functional annotation tool in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [21], we found that this set of genes is enriched in GO terms such as synapse ($p = 1.30 * 10^{-9}$), neuron development ($p = 1.60 * 10^{-5}$), exonogenesis ($p = 1.2 * 10^{-5}$) axon guidance ($p = 8.00 * 10^{-5}$) and cell adhesion ($p = 1.00 * 10^{-4}$). In particular, this set of genes is enriched in genes associated with mental retardation in the OMIM database ($p = 4.30 * 10^{-5}$), suggesting that the predicted genes to be co-regulated with the GWAS gene set may indeed be part of a network involved in g. The list of the mental retardation genes contained in our set of predicted genes is given in Table 1.

Conclusions

The D2z method is a promising tool for de novo inference of co-regulated networks of genes without reference to prior knowledge of the identity of transcription factor binding sites or transcription factors. Because of its computational efficiency and its comprehensiveness it should be generally useful as an adjunct to existing methods of transcription network inference, and especially useful for inference of networks from genome wide association studies. A major weakness to date of genome-wide association studies lies in poor statistics available for inference of underlying genomic bases when phenotypes are dependent on multiple genes, or interactions between genes. By inferring co-regulation, the D2z method serves effectively to screen, validate and extend faint signals available from genome-wide association studies. As noted in the text, other alignment-free methods could potentially be used in the same manner. Comparison of the specific methods for this purpose awaits further research.

Materials and Methods

Constructing the co-regulation graph

Given a data set containing a few human genes that are believed to be involved in the network, the promoter region of each gene in the dataset (referred to as a probe gene) is extracted, using the DBTSS database version 6 [17] including redundancies (32,122 sequences). We used the default setting in DBTSS for the length of the sequences, which is -1000 to +200. Each of these probe promoter regions is scored against the promoter regions of all other human genes in DBTSS using the D2z similarity measure [2]. The D2z measure is a normalized alignment free similarity measure that is based on the frequencies

of k-words in the sequences. The k-words represent potential binding sites of (unknown) TFs. In this analysis, the parameter k was set to be 5, representing the core length of a TFBS [2]. For each gene in the probe set we retrieve the genes associated with the top n scoring promoter sequences. For an example with 4 probe genes and $n = 5$, Figure 3a represents this step as a graph, where $g1, \ldots g4$ are the probe genes (red nodes) and each probe-node is connected by a directed edge to 5 nodes, which are the genes associated with the 5 promoter sequences that are most similar to the probe-node g by the D2z measure. The edges in the graph represent potential co-regulation between the nodes. In this example, the four probe genes produce four clusters, each with 5 arrows from the probe gene to 5 nodes. The probe genes together with the top n genes for each probe are then used to construct a co-regulation graph as follows. Common nodes between the clusters for the different probe genes are identified (Figure 3b) and only nodes that belong to more than one cluster are kept (Figure 3c). In the example, the final graph, Figure 3d, is the co-regulation graph for the probe genes $g1, \ldots g4$, where a node in the middle column is colored green if it belongs to at least three clusters and light-blue if it belongs to exactly 2 clusters.

A highly connected graph means that the probe genes have many common co-regulated genes and therefore these genes are likely to be part of a regulatory network. The example in figure 5 shows that $g1$ and $g3$ are potentially co-regulated with genes A and X (although the set of TFs involved in the co-regulations may be different in each case) and that $g2$ and $g4$ are directly co-regulated. In this example, gene A seems to be commonly co-regulated with most of the genes in the probe gene set and may be a potential hub for the network.

Assessing significance

The significance of the number of commonly co-regulated genes in the co-regulation graph was assessed using co-regulation graph produced from random genes as follows. For a co-regulation graph associated with N probes with top n hits per probe, we picked N sets, each of n random genes from DBTSS. Each of these sets corresponds to the top n genes for a probe. We constructed the co-regulation graph as before, using the $N \times n$ genes, together with the N probe nodes. The p-values were computed using 100,000 such random samples.

Acknowledgments

EJ and MRK were partially supported by the NSF grant DBI-0835718.

References

1. Davis O, Butcher LM, Docherty SJ, Meaburn EL, Curtis CJC, et al. (2010) A three-stage genome-wide association study of general cognitive ability: Hunting the small effects. Behav Genet 40: 759–767.

2. Kantorovitz MR, Robinson G, Sinha S (2007) A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23: 249–255.

3. Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, et al. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in drosophila and mouse. Cell 17: 568–579.

4. Vinga S, Almeida J (2003) Alignment-free sequence comparison - a review. Bioinformatics 19: 513–523.

5. Reinert G, Chew D, Sun F, Waterman MS (2009) Alignment-free sequence comparison (I): Statistics and power. J Comput Biol 16: 1615–1634.

6. Göke J, Schulz MH, Lasserre J, Vingron M (2012) Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. Bioinformatics : epub ahead of print.

7. Angeloni D, Lindor N, Pack S, Latif F, Wei MH, et al. (1999) CALL gene is haploinsufficient in a 3p-syndrome patient. Am J Med Genet 86: 482-485.

8. Wei M, Karavanova I, Ivanov S, Popescu N, Keck C, et al. (1998) In silico-initiated cloning and molecular characterization of a novel human member of the L1 gene family of neural cell adhesion molecules. Hum Genet 103: 355–364.

9. Angeloni D, Wei MH, Lerman MI (1999) Two single nucleotide polymorphisms (SNPs) in the CALL gene for association studies with IQ. Psychiatric Genetics 9: 165–167.

10. Montag-Sallaz M, Schachner M, Montag D (2002) Misguided axonal projections, neural cell adhesion molecule 180 mRNA upregulation, and altered behavior in mice deficient for the close homolog of L1. Molecular and Cellular Biology 22: 7967–7981.

11. Sakurai K, Migita O, Toru M, Arinami T (2002) An association between a missense polymorphism in the close homologue of L1 (CHL1, CALL) gene and schizophrenia. Molecular psychiatry 7: 412–415.

12. Frints SGM, Marynen P, Hartmann D, Fryns JP, Steyaert J, et al. (2003) CALL interrupted in a patient with non-specific mental retardation: gene dosage-dependent alteration of murine brain development and behavior. Human Molecular Genetics 12: 1463–1474.

13. Pratte M, Rougon G, Schachner M, Jamon M (2003) Mice deficient for the close homologue of the neural adhesion cell L1 (CHL1) display alterations in emotional reactivity and motor coordination. Behavioural Brain Research 147: 31–39.

14. Morellini F, Lepsveridze E, Kähler B, Dityatev A, Schachner M (2007) Reduced reactivity to novelty, impaired social behavior, and enhanced basal synaptic excitatory activity in perforant path projections to the dentate gyrus in young adult mice deficient in the neural cell adhesion molecule CHL1. Molecular and Cellular Neuroscience 34: 121–136.

15. Demyanenko GP, Siesser PF, Wright AG, Brennaman LH, Bartsch U, et al. (2011) L1 and CHL1 cooperate in thalamocortical axon targeting. Cerebral Cortex 21: 401–412.

16. Andreyeva A, Leshchyns'ka I, Knepper M, Betzel C, Redecke L, et al. (2010) CHL1 is a selective organizer of the presynaptic machinery chaperoning the SNARE complex. PLoS ONE 5: e12018.

17. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K (2008) DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res 36: D97–101.

18. Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, et al. (2009) Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N Engl J Med 360: 599–605.

19. Hamdan F, Daoud H, Piton A, Gauthier J, Dobrzeniecka S, et al. (2011) De novo SYNGAP1 mutations in nonsyndromic intellectual disability and autism. Biol Psychiatry 69: 898–901.

20. Wilczynska K, Singh S, Adams B, Bryan L, Rao R, et al. (2009) Nuclear factor I isoforms regulate gene expression during the differentiation of human neural progenitors to astrocytes. Stem Cells 27: 1173–1181.

21. Huang BDW, Lempicki R (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44–57.

Figures

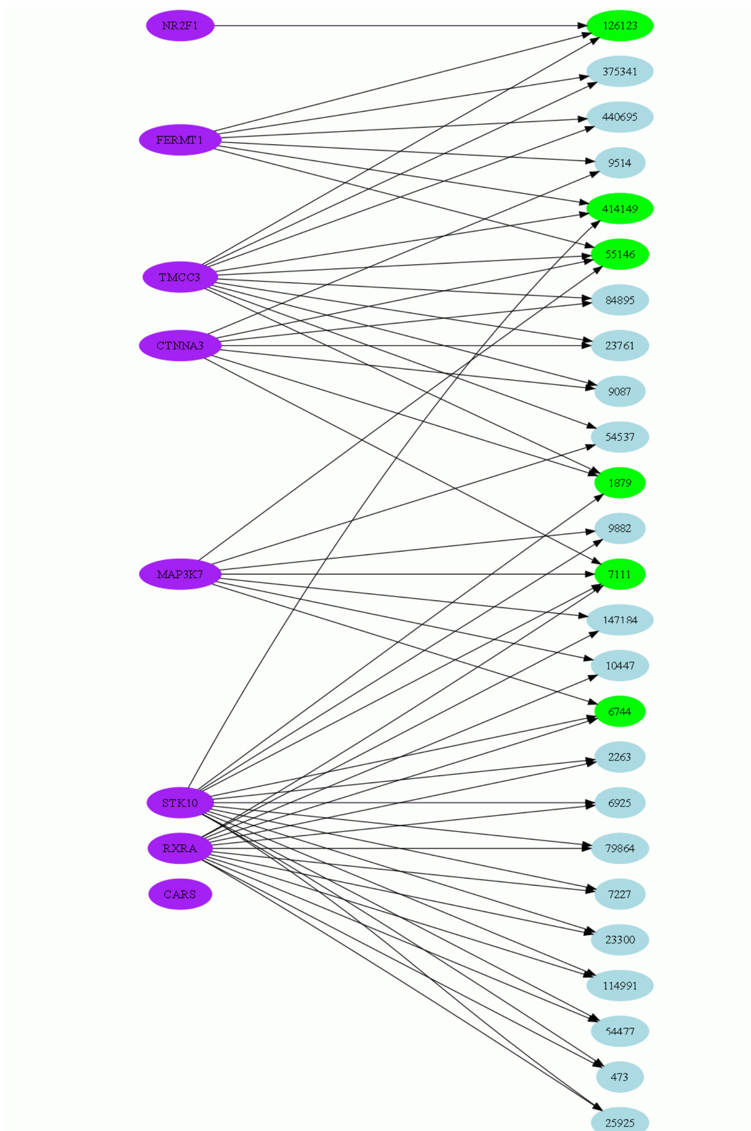

Figure 1. Co-regulation graph for the GWAS gene set. The purple nodes in the left column are the probes from the GWAS gene set. The nodes in the right column are genes that are commonly co-regulated by at least 2 probes. The green nodes are genes that are commonly co-regulated by 3 or more probes.

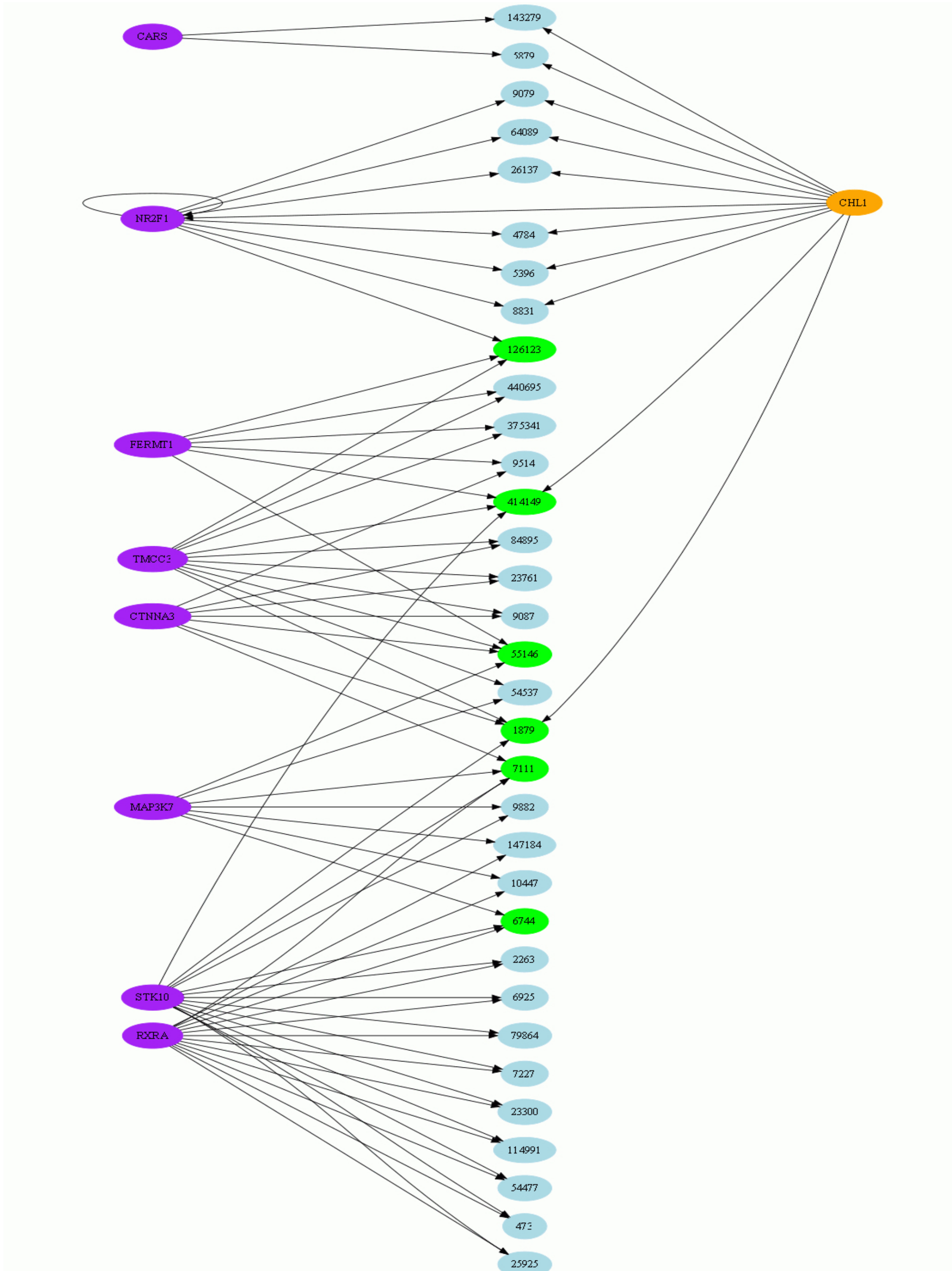

Figure 2. Co-regulation graph for the GWAS gene set and CHL1. The purple nodes in the left column are the probes from the GWAS gene set. The orange node on the right is the probe CHL1. The nodes in the middle column are genes that are commonly co-regulated by at least 2 probes. The green nodes are genes that are commonly co-regulated by 3 or more probes.

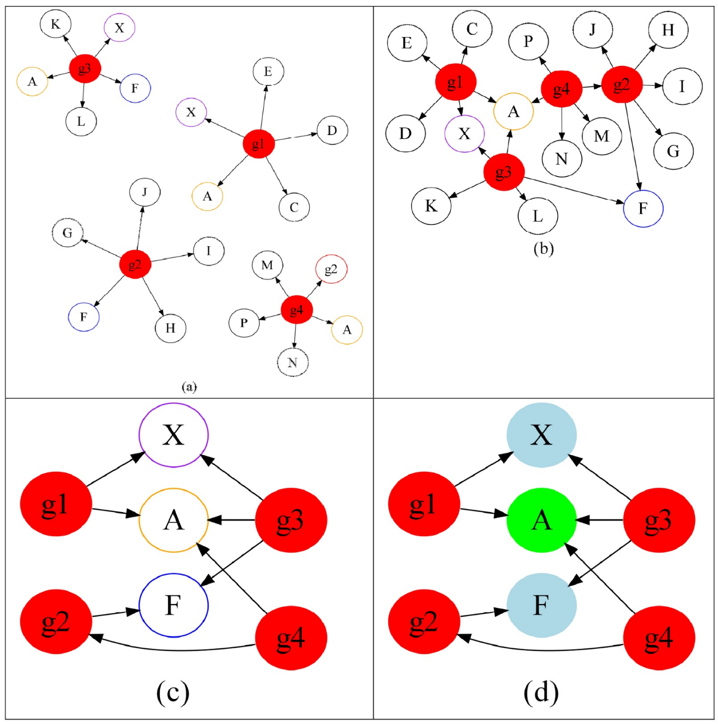

Figure 3. Constructing a co-regulation graph. In this example, the red nodes are the 4 probe genes, $g1, \ldots g4$, and the top 5 scoring genes are used for each probe. The edges in the graph represent potential co-regulation relationship between the nodes. The four probe genes produce four clusters, each with 5 arrows from the probe gene to 5 nodes (a). Common nodes between the clusters for the different probe genes are identified (b) and only nodes that belong to more than one cluster (the commonly co-regulated genes) are kept (c). The final graph (d) is the co-regulation graph for the probe genes, where the nodes in the middle column are the commonly co-regulated genes. The green nodes are genes that are commonly co-regulated by 3 or more probes and the light-blue are commonly co-regulated by 2 probes.

Tables

Table 1. Genes involved in mental retardation which are commonly co-regulated with the gene set associated with g

GENE NAME	ENTREZ GENE ID
AMMECR1	9949
L1CAM	3897
PHF6	84295
ARHGEF6	9459
ALDH3A2	224
ATRX	546
DLG3	1741
GRIK2	2898
HSD17B10	3028
MED12	9968
SYNGAP1	8831
ZEB2	9839
ZNF41	7592
PHF8	23133
PAK3	5063